Very-short-answer questions

Very-short-answer questions: reliability, discrimination and acceptability

Amir H Sam,[1, 2] Samantha M Field,[1] Carlos F Collares,[3] Cees P M van der Vleuten,[3] Val J Wass,[4] Colin Melville,[5] Joanne Harris[1] & Karim Meeran[1, 2]

CONTEXT Single-best-answer questions (SBAQs) have been widely used to test knowledge because they are easy to mark and demonstrate high reliability. However, SBAQs have been criticised for being subject to cueing.

OBJECTIVES We used a novel assessment tool that facilitates efficient marking of open-ended very-short-answer questions (VSAQs). We compared VSAQs with SBAQs with regard to reliability, discrimination and student performance, and evaluated the acceptability of VSAQs.

METHODS Medical students were randomised to sit a 60-question assessment administered in either VSAQ and then SBAQ format (Group 1, $n = 155$) or the reverse (Group 2, $n = 144$). The VSAQs were delivered on a tablet; responses were computer-marked and subsequently reviewed by two examiners. The standard error of measurement (SEM) across the ability spectrum was estimated using item response theory.

RESULTS The review of machine-marked questions took an average of 1 minute, 36 seconds per question for all students. The VSAQs had high reliability (alpha: 0.91), a significantly lower SEM than the SBAQs ($p < 0.001$) and higher mean item–total point biserial correlations ($p < 0.001$). The VSAQ scores were significantly lower than the SBAQ scores ($p < 0.001$). The difference in scores between VSAQs and SBAQs was attenuated in Group 2. Although 80.4% of students found the VSAQs more difficult, 69.2% found them more authentic.

CONCLUSIONS The VSAQ format demonstrated high reliability and discrimination and items were perceived as more authentic. The SBAQ format was associated with significant cueing. The present results suggest the VSAQ format has a higher degree of validity.

Medical Education 2018: 52: 447–455
doi: 10.1111/medu.13504

[1]Medical Education Research Unit, Imperial College London, London, UK
[2]Division of Diabetes, Endocrinology and Metabolism, Imperial College London, London, UK
[3]Department of Educational Research and Development, Maastricht University, Maastricht, the Netherlands
[4]Faculty of Medicine and Health, Keele University, Keele, UK
[5]General Medical Council, London, UK

Correspondence: Amir H Sam, 3rd Floor Hammersmith House, Hammersmith Hospital, Du Cane Road, London W12 0HS, UK.
E-mail: a.sam@imperial.ac.uk

INTRODUCTION

Multiple-choice, single-best-answer questions (SBAQs) are widely used in undergraduate and postgraduate medical assessment programmes worldwide as they tend to have high levels of reliability and can be machine-marked efficiently and accurately. However, there are long-standing concerns that multiple-choice questions (MCQs) may not provide a true reflection of knowledge as they rely on answer recognition rather than recall.[1,2] Students tend to perform better on MCQs than on open-ended, free-response, short-answer questions.[2–7] This may be attributable to the effect of cueing when candidates are presented with a list of options. A core principle of the validity of an assessment is the extent to which the test measures the competency it is supposed to measure.[8] Thus, if the aim of the assessment is to examine the student's ability to synthesise or generate rather than to recognise a correct answer, short-answer questions may provide greater validity.[9]

It is well recognised that assessment drives learning,[10,11] and that students prepare differently for examinations administered in different formats.[10,12–14] Indeed, recognition and recall require different learning operations and the distinction between them has long been recognised in cognitive psychology.[15] Learning examination technique for MCQ tests is a recognised phenomenon, which may lead to examination success at the expense of a deeper understanding of the subject being tested.[16–18] Furthermore, students demonstrate greater long-term information retention after studying for or completing short-answer questions rather than MCQs.[19–22]

A meta-analysis investigating the construct equivalence of multiple-choice and constructed response (e.g. short-answer) items showed a higher mean correlation between the two formats when stem-equivalent items were used.[23] However, studies of stem-equivalent items include assessment of topics that lend themselves to an MCQ format. In our experience, question writers can find it challenging to think of sufficient plausible distractors for some topics. The content of the assessment may also be skewed as core knowledge becomes too easy, causing question writers to resort to the testing of obscure material.[24] Short-answer questions can offer greater flexibility for question writers by allowing them to focus on common and relevant themes rather than on academic minutiae.[24,25]

Despite the potential advantages of short-answer questions, their use in large-scale assessments has been limited by feasibility as, historically, they have not been amenable to machine marking[4,26] and thus have been unable to facilitate an efficient sampling of the curriculum as a result of limitations on resources. We have previously used Microsoft Excel to mark very-short-answer questions (VSAQs).[9] In the current study, we used an online assessment management system, which allows questions to be posed on an electronic platform in a VSAQ format requiring answers of one to four words. Assessment item types can be categorised according to the degree of constraint placed on the respondent's options for answering, which range from fully constrained/selected responses (MCQs) to fully constructed responses (essays).[27] Very-short-answer questions fall into the category of 'intermediate constraint' items, which can be marked efficiently and accurately by computer using new information technologies.

The utility of an assessment method can be evaluated according to its reliability, validity, acceptability, educational impact and costs.[28] The purpose of this study was to compare VSAQs and SBAQs to assess the utility of the VSAQ format as an assessment method. Although it is difficult to link assessment formats with learning behaviour directly,[13] we assessed student opinions and the potential educational impact of VSAQs using a post-test student survey.

METHODS

Participants and assessments

This study was approved by the Medical Education Ethics Committee at Imperial College London. Ethical approval was granted to invite all medical students in Year 3 at Imperial College School of Medicine to sit a formative examination. Invitations were distributed through a faculty-sent e-mail. All students had been on one surgical and two medical attachments in Year 3. There were no other inclusion or exclusion criteria.

Medical students sat a formative examination consisting of 60 questions under examination

conditions. The clinical vignettes were constructed to allow them to be posed in both SBAQ and VSAQ formats without any change to their content (Box 1). During the construction of the VSAQs, we were able to generate a list of acceptable answers for each item within 5 minutes. Content validity was ensured by blueprinting against the Year 3 curriculum at Imperial College School of Medicine to ensure a broad sampling of relevant topics and close alignment with the syllabus. Items were written with short case descriptions and tested a range of cognitive processes, including clinical reasoning, decision making and knowledge recall. They were independently reviewed to minimise construction errors.

Box 1 Example of an item showing the five response options in a single-best-answer question (SBAQ) format (left) against the acceptable variations of the correct answer that will automatically gain a mark in the very-short-answer question (VSAQ) format (right).

A 24-year-old woman reports 2 months of lethargy, dizziness, weight loss and nausea. She has type 1 diabetes and reports erratic blood sugars and one episode of loss of consciousness. She has hyperpigmentation in her palmar creases and her oral mucosa. Her temperature is 36.8°C, pulse rate 101 bpm, blood pressure 78/61 mmHg (standing), respiratory rate 16 breaths minute^{-1} and oxygen saturation 99% breathing air. Her capillary blood glucose is 3.2 mmol/Litre.

Investigations

Sodium: 129 mmol/L (135–146)
Potassium: 5.4 mmol/L (3.4–5.0)
Urea: 7.7 mmol/L (2.5–7.8)
Creatinine: 67 μmol/L (50–95)

What is the most likely diagnosis?

SBAQ	VSAQ
A Addison's disease	*Correct answers:*
B Congenital adrenal hyperplasia	Addison's disease
C Cushing's disease	Addison's
D Hypothyroidism	Adrenal insufficiency
E SIADH	Primary adrenal insufficiency
	Hypoadrenalism

SIADH = syndrome of inappropriate antidiuretic hormone secretion, bpm = beats per minute, mmol/L = Millimoles per litre.

Examinees were randomly assigned to two groups. In Group 1, examinees were presented with 60 questions posed in VSAQ format (a 90-minute examination), immediately followed by the same 60 questions posed in SBAQ format with five options (a 60-minute examination). In Group 2, the first test consisted of SBAQs and the second of VSAQs. The students were given more time for the VSAQs to allow for the additional time required to type the answers. Students were required to complete the first test before commencing the second one, and could not return to the previous test. The VSAQs were posed using a new online examination management software (PRACTIQUE; Fry-IT Ltd, London, UK) in which students provided answers on an iPad. The SBAQ test was delivered using a traditional paper-based system with a machine-marked scoring card (MULTIQUEST; Speedwell Software, Cambridge, UK). Following their completion of the tests in both formats, students were invited to complete a feedback form in order to facilitate the evaluation of student opinions on the VSAQs.

Marking

Answers to the SBAQs were machine-marked and individual student and question performance data were exported for statistical analysis. The students' answers to the VSAQs captured by the iPad App (PRACTIQUE) were sent to a server over an encrypted connection. At the end of the examination, the server applied an automated matching algorithm using the Levenshtein distance to match each answer against preapproved acceptable answers for each question. Subsequently, all non-exact matches and match failures were reviewed by two markers to establish whether any of the non-exact matches should be disallowed and whether any of the match failures should be allowed. Similar answers were grouped in blocks to facilitate this verification process. The system applies the examiner marking judgements to all identical answers to ensure consistency and save marker time. The system also learns the new marking judgements for each question and adds this to the preapproved answer list for each question to improve the automatic marking the next time that particular question is used.

Figure 1 shows an example of the marking system. Light grey shading shows answers that are automatically marked as correct based on the preapproved answers. The unshaded answers have been marked as correct based on their similarity to preapproved answers. Answers marked as incorrect

are shown in darker shading; however, during the review process this can be over-ridden and all identical answers automatically given the same mark. The time taken by the examiners to review each item was recorded to give a measure of acceptability.

Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows Version 24.0 (IBM Corp., Armonk, NY, USA) and PRISM VERSION 5.0C (Graphpad Software, Inc., San Diego, CA, USA). The Kolmogorov–Smirnov test showed a normal distribution of all variables. Pearson's correlation coefficient was used to assess the correlation between the raw scores of the two formats. The difference in sex distribution between groups was tested using the chi-squared test. Differences in raw scores, item–total correlations and standard error of measurement (SEM) within and between groups were analysed using mixed-design analysis of variance (ANOVA) with effect sizes expressed as partial eta-squared (ηp^2). Differential item functioning for sex was assessed with XCALIBRE Version 4.2 (Assessment Systems Corp., St Paul, MN, USA) for all items using the significance of the *Z*-test based on the Mantel–Haenszel coefficient.

Students' responses to the same questions in the two formats were compared. 'Positive cueing' was defined according to the percentage of questions answered correctly in the SBAQ format and incorrectly in the VSAQ format. 'Negative cueing' was indicated when distractors caused students who provided the correct answer in the VSAQ to answer the SBAQ incorrectly.[6]

Figure 1 Example of marking of a very-short-answer question by computer. Light grey shading shows answers that are automatically assigned a mark (1.00) as they were included on the list of acceptable answers. The unshaded answer has been marked as correct based on its similarity to the acceptable answers. Answers marked by computer as incorrect are shown in darker shading; however, during the review process this can be over-ridden (e.g. 'primary hypoadrenalism'), with all identical answers automatically receiving a mark. SIADH = syndrome of inappropriate antidiuretic hormone secretion

Three-parameter logistic model analysis was carried out to include a specific parameter to estimate the probability of making a correct answer by guessing. Analysis was conducted using the R PACKAGE MIRT (R Foundation for Statistical Computing, Vienna, Austria).

In the post-test survey, students were asked to rate the following four statements on a 5-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree):

1 Questions in the single-best-answer format are easier than those in the very-short-answer format.
2 Very-short-answer questions are a better representation of how I would be expected to answer questions in clinical practice.
3 Having examinations in very-short-answer format would change my learning and revision strategy.
4 Using very-short-answer questions in assessments would help improve my preparation for clinical practice.

The questionnaire also provided a space for students to write comments about the use of VSAQs in medical school assessments. NVIVO Version 10 (QSR International Pty Ltd, Melbourne, VIC, Australia) was used to identify themes in the students' free-text responses.

RESULTS

Of 340 students in Year 3 at Imperial College School of Medicine, 302 students (88.8%) sat the formative examination and 299 (99.0%) completed both parts. The sex distribution was similar in both groups (female : male ratio: Group 1: 72 : 83; Group 2: 63 : 81; $p = 0.64$). Among the tests in both groups, only one item (a VSAQ in Group 1) had significant differential item functioning for sex, with bias against males ($p = 0.04$).

There was a significant positive correlation between the two formats ($p < 0.001$, $r = 0.83$). In view of the high reliability and identical content of the tests, correction for attenuation was not performed as this was likely to overestimate the correlation.

Acceptability

The VSAQs were reviewed by two examiners. Based on the preloaded acceptable answers, the system was able to identify 80.2% of correct answers prior to review, which was instrumental in allowing efficient marking. The remainder of correct answers were marked during the review process. Of answers marked 'correct' by the system, 0.2% were deemed to be 'incorrect' on review (because a spelling error significantly changed the meaning of the answer). The total time taken to review the machine-marked answers to all 60 VSAQs for all 299 students was 95 minutes, 51 seconds. The average time spent by examiners on reviewing the answers to each question for all 299 students was 1 minute, 36 seconds (standard deviation: 1 minute, 2 seconds). The marking system allowed for multiple correct answers in addition to trivial differences in spelling or terminology, and in 8.3% (5/60) of questions at least one student offered an alternative answer to the question that was judged to be correct following the review by examiners.

Effect of cueing

The raw scores for each test are shown in Table 1. The raw scores for VSAQs and SBAQs in Group 1 were 52.4% and 68.2%, respectively. In Group 2, the raw scores for VSAQs and SBAQs were 65.7% and 69.7%, respectively. There was a significant difference in raw scores between item types with a large effect size ($F_{(1,297)}$ = 384 339, $p < 0.001$, $\eta p^2 = 0.56$). There were also significant differences in the interaction between item type and group ($F_{(1,297)}$ = 136 343, $p < 0.001$, $\eta p^2 = 0.32$) and between groups ($F_{(1,297)}$ = 19 854, $p < 0.001$, $\eta p^2 = 0.06$). The difference between the two groups in VSAQ and SBAQ scores is likely to reflect the cueing effect associated with the fact that Group 2 students saw the answer options in the SBAQs before they answered the VSAQs. Positive cueing, whereby students answered the SBAQ correctly and the equivalent VSAQ incorrectly, was seen in 19.2% of items in Group 1 and 7.5% of items in Group 2. Negative cueing, whereby students answered the VSAQ correctly and the equivalent SBAQ incorrectly, was in seen in 3.5% of items in Group 1 and 3.5% of items in Group 2. On an item level, positive cueing occurred for every item (100%) and negative cueing occurred in 50 of 60 (83.3%) items. On four items students performed better in the VSAQ than the SBAQ format, possibly because the answer options distracted the students. Notably, in 20.0% of items, over 30.0% of students in Group 1 obtained the correct answer only in the SBAQ format. For example, whereas only 13.5% of students in Group 1 were able to generate a diagnosis of erythema multiforme in the VSAQ, 67.7% were able to select the correct answer in the SBAQ.

Reliability and discrimination

Table 1 shows reliability (Cronbach's alpha) and SEM values for the VSAQ and SBAQ tests in both groups. Tests using the SBAQ format had Cronbach's alpha values of 0.84 and 0.85 in Groups 1 and 2, respectively. Tests using the VSAQ format had a Cronbach's alpha value of 0.91 in both Groups 1 and 2.

We also used the three-parameter logistic model to estimate SEMs for the two tests in both groups across the ability spectrum (Fig. 2). Items formatted

Table 1 Mean ± standard deviation (SD) raw scores, Cronbach's alpha, standard error of measurement (SEM) and mean item–total score point-biserial correlations for the very-short-answer question (VSAQ) and single-best-answer question (SBAQ) tests in Groups 1 and 2

	Group 1 (*n* = 155)		**Group 2 (*n* = 144)**	
	VSAQ	**SBAQ**	**SBAQ**	**VSAQ**
Mean ± SD score, %	52.4 ± 17.4%	68.2 ± 12.5%	69.7 ± 12.9%	65.7 ± 16.5%
Cronbach's alpha	0.91	0.84	0.85	0.91
SEM	5.24	5.03	4.97	5.09
Mean item–total score point-biserial correlation	0.36	0.26	0.27	0.35

as VSAQs had significantly lower SEM values ($F_{(1,297)}$ = 213 782, p < 0.001, ηp^2 = 0.42). The effects of group and the interaction between item format and group were not significant (p = 0.23 and p = 0.97, respectively).

Figure 3 shows individual estimates for information according to theta ability estimates and item type and group combination. Items formatted as VSAQs had significantly higher information with a large effect size ($F_{(1,297)}$ = 311 998, p < 0.001, ηp^2 = 0.51). There was no significant interaction between item format and group ($F_{(1,297)}$ = 3584, p = 0.06, ηp^2 = 0.01) and the group effect size was small ($F_{(1,297)}$ = 4742, p = 0.03, ηp^2 = 0.02).

Mean item–total score point-biserial correlations for VSAQs were 0.36 and 0.35 in Groups 1 and 2, respectively. Mean item–total score point-biserial correlations for SBAQs were 0.26 and 0.27 in Groups 1 and 2, respectively. Items formatted as VSAQs had significantly higher item–total correlations ($F_{(1,118)}$ = 89 235, p < 0.001, ηp^2 = 0.43). The effects of group and the interaction between format and group were not significant (p = 0.81 and p = 0.30, respectively).

Potential impact on learning behaviour

Figure 4 shows the percentage of students selecting each point on the Likert scale for each statement. A total of 80.4% of students agreed or strongly agreed that SBAQs were easier than VSAQs. With regard to authenticity, 69.2% agreed or strongly agreed that VSAQs were more representative of how they would be expected to answer questions in clinical practice. Almost half the cohort (49.3%) agreed or strongly agreed that having VSAQs in summative examinations would change their revision and learning strategies and would help improve their preparation for clinical practice.

Overall, 30.9%of students who thought SBAQs were easier commented that this was attributable to the presence of options. A total of 56.1% of free-text comments regarding authenticity referred to how, in practice, students will be expected to recall information without options. Of comments on making changes in revision strategies, 69.7% indicated recognition of a need for more emphasis on thoroughness and 9.1% indicated a need to spend more time learning spelling.

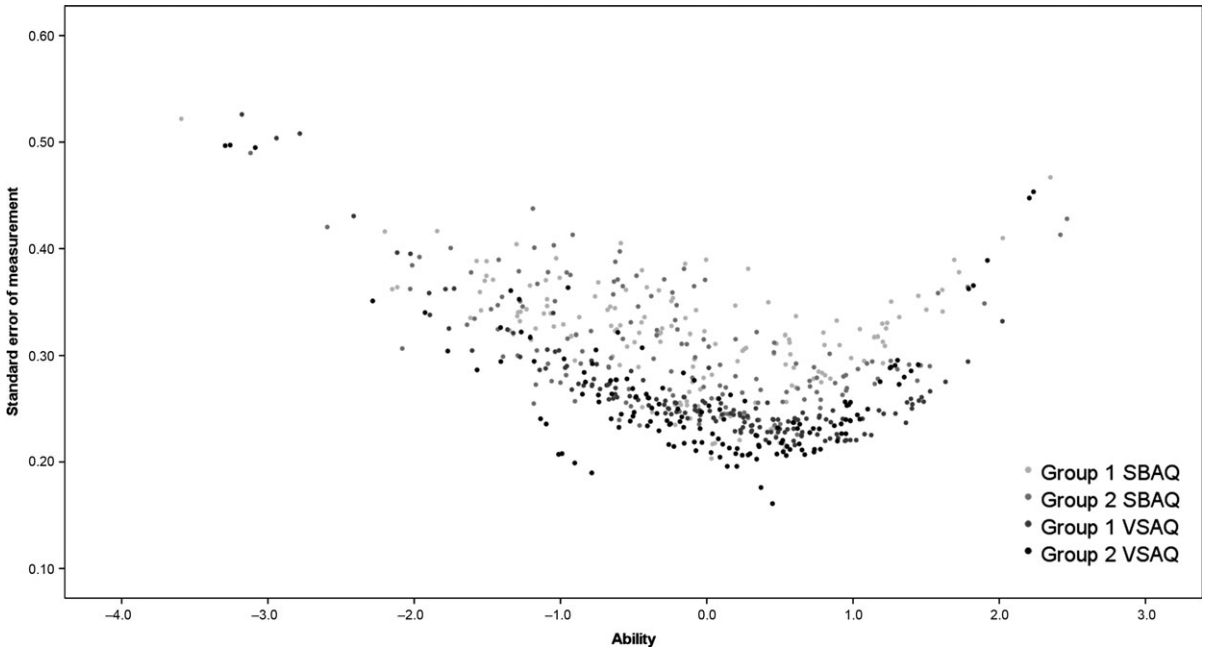

Figure 2 Individual estimates for the standard error of measurement according to theta ability estimates, and item type (single-best-answer question [SBAQ]; very-short-answer question [VSAQ]) and group combination

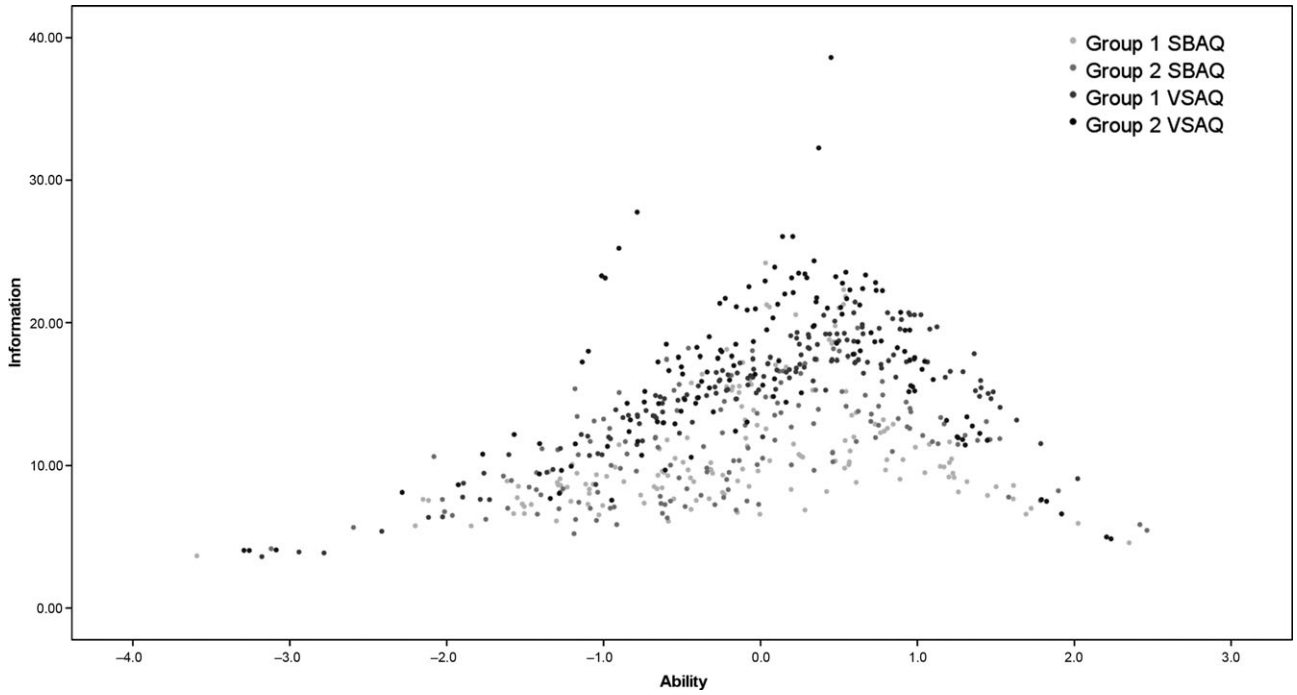

Figure 3 Individual estimates for information according to theta ability estimates, and item type (single-best-answer question [SBAQ]; very-short-answer question [VSAQ]) and group combination

DISCUSSION

The present results indicate that VSAQs have advantages over SBAQs in terms of reliability and discrimination. The cueing effect associated with SBAQs was apparent in the analysis of the scores for the VSAQs and SBAQs in the two groups. The difference in scores between the two formats was attenuated when the students saw the SBAQs first and is likely to be attributable to the cueing effect of the options. Therefore, VSAQs have higher validity in testing the ability to arrive at a correct answer without cueing or guessing.

Another potential limitation of SBAQs is the implication that there is one best answer for any question, which discourages question writing in

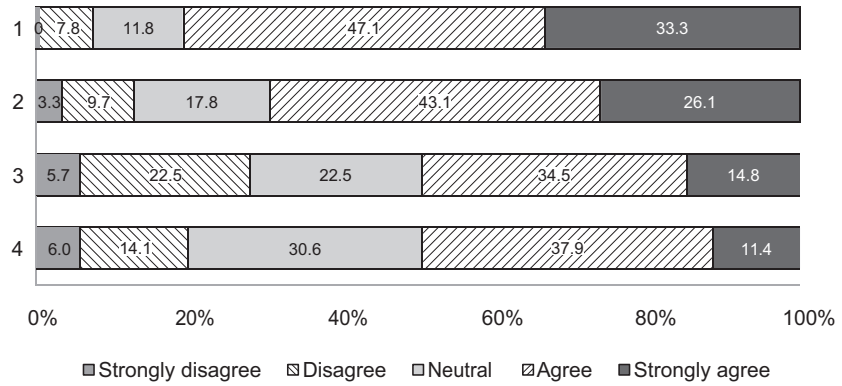

Figure 4 Students were asked to rate their agreement with each of four statements using a 5-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree). The percentages of students selecting each point on the Likert scale are shown for each of four statements: 1 = Questions in the single-best-answer format are easier than those in the very-short-answer format; 2 = Very-short-answer questions are a better representation of how I would be expected to answer questions in clinical practice; 3 = Having examinations in very-short-answer format would change my learning and revision strategy, and 4 = Using very-short-answer questions in assessments would help improve my preparation for clinical practice

areas of medicine in which there may be multiple defensible answers.[1,3] The VSAQ format allows students to offer alternative answers that may be as good or may be second-option alternatives. Allowing students to demonstrate the scope of their knowledge improves the validity of the assessment.

Almost 70% of students thought VSAQs better represented how they would be expected to answer questions in real-life clinical practice, which suggests that VSAQs have greater authenticity. More authentic examinations both promote deeper learning methods and increase student motivation.[29] Furthermore, a major component of learning from assessment is the quality of the feedback provided.[8,11] Using items formatted as VSAQs can offer opportunities to provide more specific and detailed feedback based on the diverse range of incorrect answers proposed by students. These can be addressed or targeted by curriculum developers.

A major limitation to the widespread use of VSAQs may be their acceptability in terms of resources. The value of introducing any novel assessment method should be weighed against the extra time and use of resources incurred. Machine-marked VSAQs should be reviewed by subject matter experts. With the assessment software used in this study, the total time taken to review the machine-marked answers to all 60 VSAQs for 299 students was under 2 hours. Items that took longer to mark had a higher number of permutations of the correct answers, which made it difficult to create a comprehensive answer key (e.g. ultrasound left leg, Doppler USS LL, left leg US). However, every time an unforeseen answer is marked as 'correct' by the examiner, the system will learn the variation. Therefore, in future uses of the question, the fact that the answers will already be present in the system will make marking more efficient. With advances in computational linguistics and machine learning across educational fields,[30,31] it is likely that the speed and reliability of marking short-answer questions will continue to improve.

The limitations of this study include its sample size and the inclusion of students from a single centre. Although the students were randomly assigned to the two groups and there were no differences in sex distribution, the possibility of differences in other characteristics cannot be excluded. For example, we did not have access to data on the cultural backgrounds of students for inclusion in the differential item functioning. Only one VSAQ showed differential item functioning with a bias against males. Interestingly, this item was based on the presentation of urinary tract infection in a young female. Another limitation of the study concerns the lack of data on the clinical experience and anticipated specialty of those who agreed VSAQs were more representative of real-life practice. Furthermore, there was no external validation measure to assess and compare students' competence. Future research should investigate the utility of VSAQs in multicentre studies involving larger numbers of students. It would also be interesting to examine the effects of cueing in candidates with varying levels of expertise.

Given their high correlation with short-answer questions, SBAQs are widely used as an efficient assessment tool across medical education as a proxy for assessing applied knowledge. However, high correlations between assessments do not necessarily imply that they test the same cognitive facility. Indeed, answer generation rather than recognition is tested in short-answer questions.[2,4,32] Items formatted as VSAQs have levels of efficiency and acceptability that approach those of SBAQs. Furthermore, compared with the SBAQ format, VSAQs demonstrate higher reliability, discrimination and authenticity. The results of this study demonstrate the utility of the VSAQ-based test as an assessment instrument that has the potential to improve existing assessment programmes.

Contributors: All authors contributed to the conception and design of the work, the analysis and interpretation of data, and the drafting and critical revision of the paper. All authors approved the final manuscript for submission.
Acknowledgements: The authors thank Keean Schupke, Fry-IT Ltd, London, UK, for his help with the provision of the PRACTIQUE software.
Funding: None.
Conflicts of interest: None.
Ethical approval: The study protocol was approved by the Medical Education Ethics Committee at Imperial College London.

REFERENCES

1 Elstein AS. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. *Acad Med* 1993;**68** (4):244–9.
2 Veloski JJ, Rabinowitz HK, Robeson MR, Young PR. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Acad Med* 1999;**74** (5):539–46.

3 Shaibah HS, van der Vleuten CP. The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy. *Anat Sci Educ* 2013;**6**:149–56.
4 Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Med Educ* 1979;**13** (4):263–8.
5 Desjardins I, Touchie C, Pugh D, Wood TJ, Humphrey-Murto S. The impact of cueing on written examinations of clinical decision making: a case study. *Med Educ* 2014;**48** (3):255–61.
6 Schuwirth LWT, van der Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. *Med Educ* 1996;**30** (1):44–9.
7 Schuwirth LWT, van der Vleuten CPM, Stoffers HEJH, Peperkamp AGW. Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. *Med Educ* 1996;**30** (1):50–5.
8 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.
9 Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Med Educ* 2016; **16** (1):266.
10 Epstein RM. Assessment in medical education. *N Engl J Med* 2007;**356** (4):387–96.
11 Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;**357** (9260):945–9.
12 Cilliers FJ, Schuwirth LW, van der Vleuten CP. A model of the pre-assessment learning effects of assessment is operational in an undergraduate clinical context. *BMC Med Educ* 2012;**12**:9.
13 Al-Kadri HM, Al-Moamary MS, Roberts C, van der Vleuten CP. Exploring assessment factors contributing to students' study strategies: literature review. *Med Teach* 2012;**34** (suppl 1):42–50.
14 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;**17** (3):165–71.
15 Eagle M, Leiter E. Recall and recognition in intentional and incidental learning. *J Exp Psychol* 1964;**68**:58–63.
16 McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach* 2004;**26** (8):709–12.
17 Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. *Med Educ* 1986;**20** (3):162–75.
18 Willing S, Ostapczuk M, Musch J. Do sequentially presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Adv Health Sci Educ Theory Pract* 2015; **20** (1):247–63.
19 McConnell MM, St-Onge C, Young ME. The benefits of testing for learning on later performance. *Adv Health Sci Educ Theory Pract* 2015;**20** (2):305–20.
20 Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. *Med Educ* 2008;**42** (10):959–66.
21 Wood T. Assessment not only drives learning, it may also help learning. *Med Educ* 2009;**43** (1):5–6.
22 McDaniel MA, Roediger HL III, McDermott KB. Generalising test-enhanced learning from the laboratory to the classroom. *Psychon Bull Rev* 2007; **14** (2):200–6.
23 Rodriguez MC. Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *J Educ Measure* 2003;**40** (2):163–84.
24 Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Hum Pathol* 1997;**28** (5):526–32.
25 Damjanov I, Fenderson BA, Veloski JJ, Rubin E. Testing of medical students with open-ended, uncued questions. *Hum Pathol* 1995;**26** (4):362–5.
26 Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. *Teach Learn Med* 1993;**5** (2):107–15.
27 Scalise K, Gifford B. Computer-based assessment in e-learning: a framework for constructing 'intermediate constraint' questions and tasks for technology platforms. *J Technol Learn Assess* 2006;**4** (6):1–44.
28 van der Vleuten CP. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996; **1** (1):41–67.
29 Gulikers JTM, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. *Educ Technol Res Dev* 2004;**52** (3):67–86.
30 Burrows S, Gurevych I, Stein B. The eras and trends of automatic short answer grading. *Int J Artif Intell Educ* 2015;**25** (1):60–117.
31 Pulman SG, Sukkarieh JZ. Automatic short answer marking. Proceedings of the Second Workshop on Building Educational Applications Using NLP, 29 June 2005, Ann Arbor, MI;9–16.
32 Ozuru Y, Briner S, Kurby CA, McNamara DS. Comparing comprehension measured by multiple-choice and open-ended questions. *Can J Exp Psychol* 2013;**67** (3):215–27.

Received 21 June 2017; editorial comments to author 25 July 2017; accepted for publication 7 November 2017